BMC Genomics

RESEARCH Open Access

Contrastive sequence signatures between the both sides of a recombination spot reveal an adaptation at *PPARD* locus from standing variation for pleiotropy since out-of-Africa dispersal

Weihua Shou[1,2*], Chenhui Zhang[2], Ying Wang[2], Haifeng Wang[2], Lei Guo[1], Li Li[1], Tiesong Zhang[1*], Wei Huang[2] and Jinxiu Shi[2*]

Abstract

Background Drug metabolism and transporter genes are a specialized class of genes involved in absorption, distribution, metabolism and excretion. They easily present distinct genetic population differentiation and are vulnerable to natural selection.

Results We initiated a study using a special panel of informative genetic markers in such genes and dissected the genetic structure in representative Chinese and worldwide populations. A distinctive sub-population stratification was discovered in extensive Eurasians and resulted from divergence at the *PPARD* locus. The contrastive sequence signatures between the both sides of a recombination spot prove a selective sweep on this locus for genetic hitchhiking effect. A genealogy-based framework demonstrates the positive selection acting from standing variation exerted a moderate pressure in Eurasians, and drove the adaptive allele up to a high frequency. The timing and tempo estimations for the genetic adaptation indicate its onset coincided with the early out-of-Africa migration of modern humans and it lasted over a prolonged evolutionary history. A phenome-wide association analysis reveals an extended *cis*-regulation on the local gene expression and the pleiotropy implicated in a variety of complex traits. The colocalization analyses between the genetic associations from *cis*-acting gene expression and complex traits signify the most likely selective pressure from physical capacity, energy metabolism, and immune-related involvement, and provide prioritization for the effective genes and casual variants.

*Correspondence:
Weihua Shou
shouwh_chgc@163.com
Tiesong Zhang
zts68420@sina.com
Jinxiu Shi
shijx@chgc.sh.cn

Full list of author information is available at the end of the article

Conclusions This work has laid a foundation for following efforts to make full sense of the biological mechanisms underlying the genetic adaptation.

Keywords Drug gene, Genetic stratification, Positive selection, Evolutionary history, Selective pressure, Causal prioritization

Background

Over the past few years, intensive efforts have been made to elucidate more precise genetic architecture throughout vast Chinese populations with ethnic diversity via scrutinizing genome-wide variations from early high-density microarray genotyping [1–2] and recent whole genome sequencing [3–4]. The genetic markers are usually more preferred to enable discriminating subtle population differentiation, regardless of their biological functionality. In concomitant analyses for natural selection, a series of genetic loci have been identified under positive selection, especially recent hard sweep [5]. Exploring these evolutionary profiles and their phenotypic implications has been essential for better understanding the human genome. We anticipated that a detailed survey for the genes involving certain biological functions would provide novel insights into what has not been achieved in previous works.

Drug metabolism and transporter genes are a specialized class of genes involved in absorption, distribution, metabolism and excretion (ADME), and significantly contribute to human variability in drug response [6–8]. The PharmacoGenetics for Every Nation Initiative aimed to establish a global genotype-guided knowledge resource to formulate individualized therapy for drug efficacy and safety [9]. An exemplified application is the major genetic determinants of warfarin dosing strategy for optimal anticoagulation [10–11] within a narrow therapeutic range for patients with wide dosage requirements. Moreover, the spectrum of ADME genetic variants often reflect easily discernible population heterogeneity. The differentiation is not only resulted from random genetic drift, human demographic activity and sociocultural transition, but also usual susceptibility to natural selection due to their involvement in crucial biological processes [12]. Despite constantly receiving unusual attention, such genes remain ambiguous in many aspects of human genetics.

We initiated the present study using a special panel of informative genetic markers in the ADME genes, followed by a series of progressive analyses for genetic population differentiation, sequence signatures of genetic adaptation, selective history inference, potential selective pressure and prioritization of casual genes and variants.

Results

Population structure in ADME genes

We carried out a popular workflow of population genetics to characterize the genetic architecture of the ADME genes among the representative Chinese populations (Table S1). Our results from a particular gene class reiterate a well-recognized genetic structure (Fig. 1) among Chinese populations. Three major groups, namely the northern minority (Uygur, Mongolian and Tibetan), Han Chinese (Han-GD, Han-SH and Han-SD) and southern minority (Lic, Zhuang and Miao) groups, primarily reflect genetic changes in a geographical south-north cline (Fig. 1a and c). The divergence between the southeastern and the northwestern minority ethnic groups is more striking. The northern populations diverge from one another and the other populations (Fig. 1b-c), with their distinct genetic diversity representing part of an east-west stratification [13–15]. The change in the ancestry composition of each STRUCTURE analysis is more significant in the northern populations (Fig. S1). The variance percentages among groups turn higher when separately grouping the three northern populations in the Analysis of Molecular Variance (AMOVA) analyses, and the grouping in agreement with their language phylum yields the highest among-group variance (Table S2). The differentiation is sophisticated within the Han Chinese and southern minority populations. Han-GD shows a closer relationship to Lic and Zhuang than to Han-SH and Han SD (Fig. 1c) because of their genetic make-up resemblance in the STRUCTURE Clusters (Fig. 1b). The variance percentage among groups is larger when grouping Han-GD with Lic and Zhuang than when grouping it with Miao in the AMOVA. STRUCTURE (Fig. S1) and AMOVA (Table S2) equally reveal Miao seems more deviated, although it belongs to one of the southern populations. This specialized combination of genetic markers is less able to clearly separate the populations, especially the southern populations due to their closer relationships resulted from intricate demographic history. It was still challenging, yet improved, even when using the whole genome unbiased data [4, 16]. All Clusters in the STRUCTURE analyses are shared between the Han Chinese and the southern populations, except for a discernible variation in each Cluster at the population level (Fig. S1). Most importantly, the STRUCTURE characterization of genetic architecture indicates a notable genetic stratification between individuals, even within the same population. Consistently, a plausible binary clustering

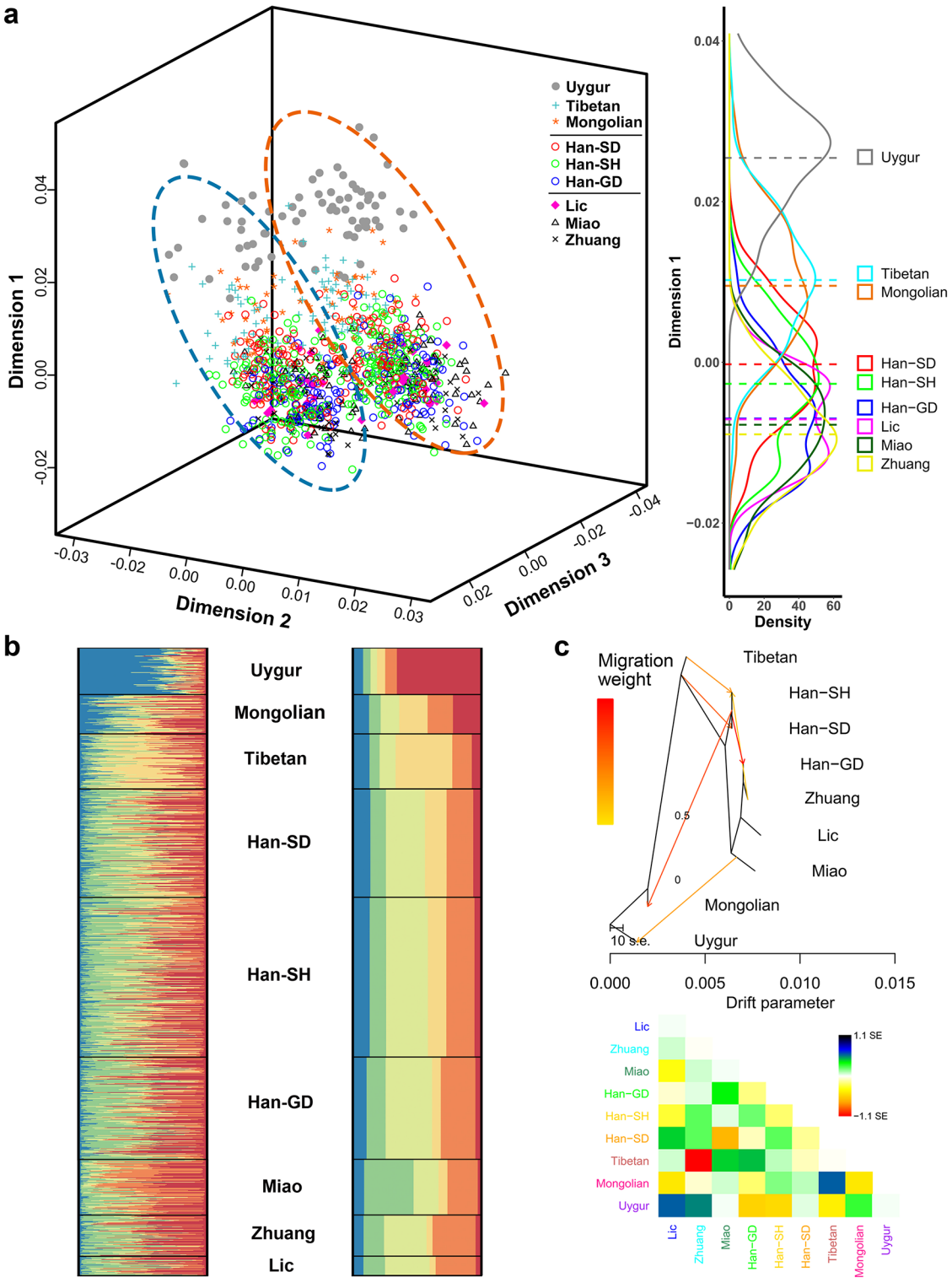


Fig. 1 Population structure in the ADME genes. **a** The MDS plot of the Chinese populations in three dimensions on the basis of pairwise IBS matrix and the distribution of each population in the first dimension. The sub-population clustering is marked by dashed circles in different colors. **b** The STRUCTURE result at K = 6 exhibits the varying proportions of individual Clusters. The left and right parts depict for the individual levels and the population levels, respectively. **c** The maximum likelihood tree of TreeMix for the populations with the migration weight and residuals from the optimal model fit. Uygur was used as the outgroup in the TreeMix modeling. The x-axis represents the estimated relative genetic drift. The migration edge was set at 6 after an evaluation (Fig. S2-S3)

pattern is appreciated in the Multi-dimensional Scaling (MDS) analysis (Fig. 1a). The proportions of certain Clusters anomalously fluctuate between individuals throughout the populations in all the STRUCTURE analyses at the individual level (Fig. S1). Thus, we recognized a sub-population structure that has never been documented.

Pairwise F_{ST} values between populations were estimated for each genetic marker to measure the divergence of allele frequency. The extent of divergence corresponds to the population structure demonstrated in the aforementioned analyses. High values are easily observed between the northern minority populations and the others (Fig. S4; Table S3). The top percentiles (99.9th, 99th, 98th and 95th) of the pairwise population F_{ST} values exhibit the lowest differentiation among the Han Chinese population pairs. The 99.9th percentile is F_{ST}=0.308, and the 99th percentile is F_{ST}=0.14 for the all F_{ST} value distribution. If a SNP (single nucleotide polymorphism) has F_{ST}≥0.3 or F_{ST}≥0.15, it is considered as the most or very highly differentiated, respectively (Tables S4-S5). The 71 SNPs with the highest F_{ST} values are associated with 23 genes. Of them, a few are remarkable for significance in clinical pharmacogenetics [17], like *CYP2C19*, *CYP2D6* and *VKORC1*, for being the target of natural selection, like *ADH1B* [18] and *NAT2* [19], for being involved in disease pathogenesis, like *ABCG2* [20], *ALB* [21] and *CYP11B1* [22], and for a dual role in transmembrane transportation and viral entry, like *SLC10A1* [23]. The well-known *ADH1B*48His* (rs1229984) in *ADH1B* is strikingly differentiated between Tibetan and other populations, with the highest F_{ST}=0.646 between Tibetan and Lic. An adaptation to the subsistence lifestyle following the advent of Neolithic agriculture has been broadly accepted as the evolutionary drive [18]. Accordingly, genetic divergence seems easier to occur in the ADME genes.

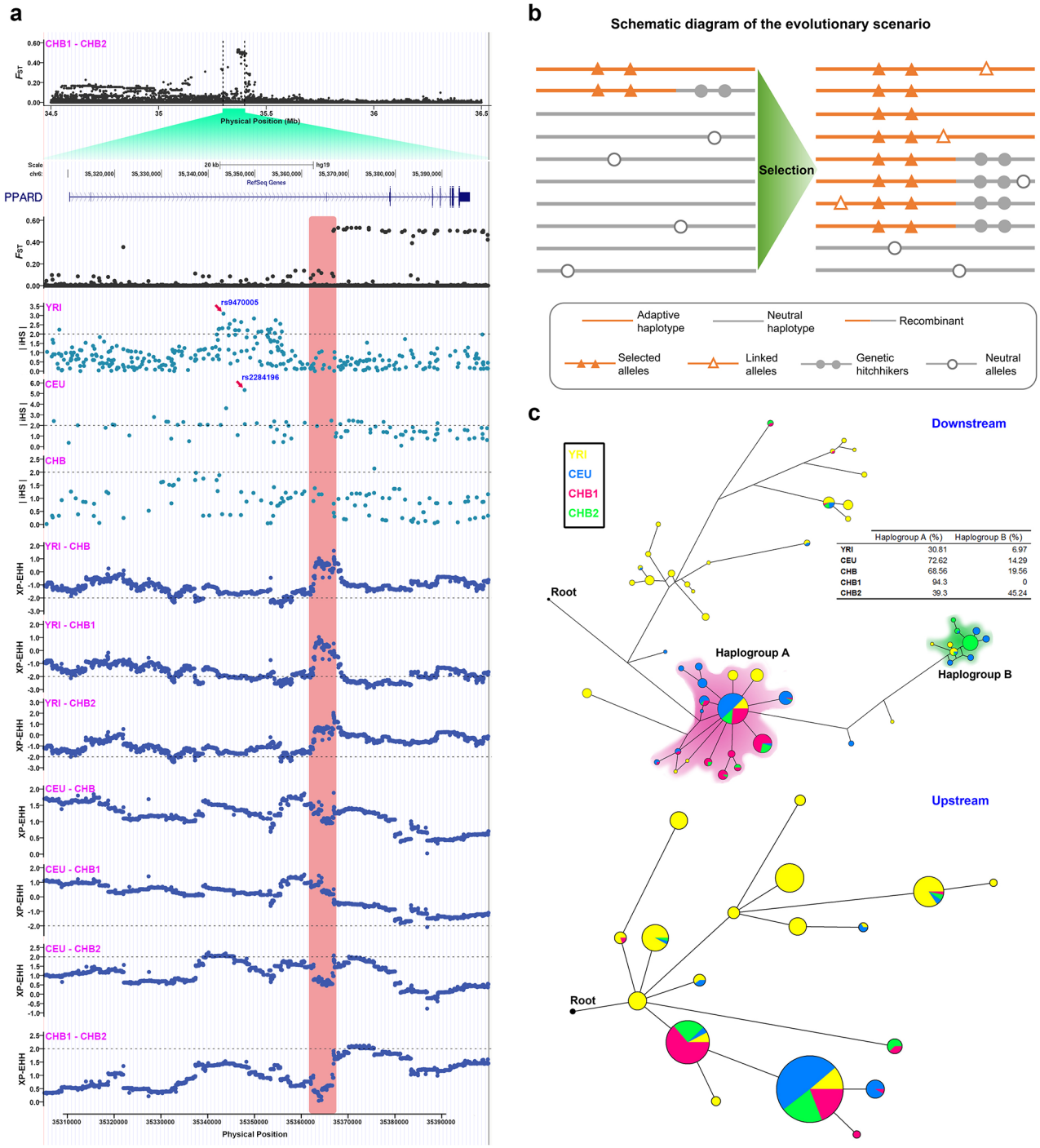
Using a particular class of genetic markers in the ADME genes, our results recapitulated the general population structure in Chinese populations, found out the genetic differentiation in a part of potential markers, and discovered a paradoxical genetic split between the individuals of each population (Fig. 1a). Thus, we sought to examine the factors gave rise to this extensive and notable genetic distinction, which had never been disclosed in previous works.

Distinctive sub-population stratification

The conventional analyses of population genetics added an unexpected insight to the results that largely align with existing knowledge on the genetic structure in East Asians. A distinctive subdivision between the individuals across all the populations is worthwhile to explore its root (Fig. 1a). Previous studies adopted a linkage disequilibrium (LD)-based thinned set of genome-wide polymorphisms as a balanced strategy to dissect population structure in case of uncertain genetic bias [24]. However, the potential ability may be attenuated to identify subtle differentiations at specific loci. Subsequently, we strove to ascertain whether this disparity is owing to an unusual but genuine population structure or a spurious chance arising from our biased analysis and even genotyping errors. Additionally, we proceeded with inquiry into the underlying cause and why it occurs within populations, if the sub-population stratification does exist.

We analyzed the population structure in a more global context. Data from the 1000 Genomes Project (1KGP) [25] shared with our DMET genotype data were exploited. Additional 14 populations in 1KGP, including Han Chinese in Beijing, China (CHB), Han Chinese South (CHS) and Japanese in Tokyo, Japan (JPT) of East Asian ancestry, British from England and Scotland (GBR), Finnish in Finland (FIN), Iberian Populations in Spain (IBS), Toscani in Italia (TSI) and Utah residents (CEPH) with Northern and Western European ancestry (CEU) of European ancestry, Colombian in Medellín, Colombia (CLM), Mexican Ancestry in Los Angeles CA, USA (MXL), and Peruvian in Lima, Peru (PEL) of Latin American ancestry, and African Ancestry in Southwest USA (ASW), Luhya in Webuye, Kenya (LWK) and Yoruba in Ibadan, Nigeria (YRI) of African ancestry, were integrated into an MDS analysis with the present samples (Fig. S5). All subjects were clustered into three major groups representing East Asians, Europeans and Latin Americans, and Africans. The Latin Americans clustering is close to that of Europeans because of a considerable contribution from European ancestry into Latin American populations [26]. An intermediate distribution of Uygur between East Asians and Europeans abide by our expectations. Whereas, the 1KGP population clustering consistently confirms what was manifested in our samples. An appreciable genetic stratification remains in the East Asian populations and even extends to the European and Latin American populations. This unusual sub-population structure barely arose from genotyping and analytical mistakes, but suggested an unidentified genetic differentiation across Eurasians.

We took a further look into the sub-population structure for its cause. The Han Chinese and southern minority populations are easily allocated into two sub-populations based on their clustering. Pairwise F_{ST} of individual SNPs was calculated for all sub-populations (Fig. S6). The analysis before enabled to filter high F_{ST} values due to divergence between different ethnic populations. Hence, an identical SNP set with prominent F_{ST} values was generated from all pairs of the different clustered subpopulations. These SNPs exactly pinpoint to a well-known gene, peroxisome proliferator-activated receptor delta (*PPARD*), a master regulator in the PPAR



	Haplogroup A (%)	Haplogroup B (%)
YRI	30.81	6.97
CEU	72.62	14.29
CHB	68.56	19.56
CHB1	94.3	0
CHB2	39.3	45.24

Fig. 2 (See legend on next page.)

(See figure on previous page.)

Fig. 2 The sequence signatures of positive selection at *PPARD*. **a** Contrastive sequence variation profiles between the both sides of a recombination point at *PPARD*. The F_{ST}, iHS and XP-EHH statistics were calculated using the 1KGP phased genotypic data. A vertical block in light brown corresponds to the recombination spot (Fig. S7). The upper part depicts the F_{ST} values between the two CHB subpopulations surrounding the *PPARD* region. The middle part depicts the regional plots of iHS for YRI, CEU and CHB. The lower part indicates the regional plots of XP-EHH estimated between different pairs of populations. A cutoff of ±2 was adopted and denoted in the dashed lines in the plots. **b** A schematic diagram of the proposed evolutionary scenario. The orange segments represent the adaptive haplotypes carrying the selected allele, the grey segments the neutral haplotypes. A neutral haplotype links to an adaptive haplotype after a recombination. The neutral alleles on the recombinant begin to increase as the adaptive haplotypes expand in a population under selection. The neutral alleles on the recombinants become genetic hitchhikers. **c** Median-joining networks for the *PPARD* haplotypes of the separate regions in major continental populations. The networks were constructed from 530 haplotypes composed of common SNPs in each population. The ancestral allele haplotypes were used as outgroup in analysis. Each node represents a haplotype and its size is proportional to the haplotype frequency. The upper part depicts the networks constructed using 156 SNPs in the downstream region of the recombination spot. Two haplotype clusters were defined as two haplogroups in different color shadows. The distribution of the two haplogroups was estimated. The lower part portrays the networks constructed using 27 SNPs in the upstream region

sub-family of nuclear hormone receptors [27–29]. Its hallmark capabilities have been documented in many biological processes, such as immunity, energy and lipid metabolism, angiogenesis, cell death, cardiovascular health, and tumor growth, invasion, and metastasis. The activation of PPARδ plays a protective role in metabolic and inflammatory conditions, promising a target for treating metabolic and immune-related diseases. A fundamental resource of genome-wide scans for recent positive selection [30], which developed the integrated haplotype score (iHS) measuring the signature of positive natural selection, has provided suggestive evidence to a possible selective effect on *PPARD* (Table S6).

Natural selection usually leads to genetic differentiation among populations from different geographical regions with changed ethnic or cultural backgrounds, and natural habitats. Selective force appears ironic in explaining the striking subpopulation divergence. It became of interest to illuminate the *PPARD* mystery.

Selection driving the differentiation at *PPARD*

We harnessed the model populations from the 1KGP data to elucidate the evolution at the *PPARD* locus. The CHB participants were assigned into two subpopulations, designated as CHB1 and CHB2, in accordance to the above clustering (Fig. S5). F_{ST} values were calculated between CHB1 and CHB2 using the 1KGP phased genotypic data on a larger genomic scale. A focus on the 2 Mb syntenic region centered on the *PPARD* gene shows that the highest F_{ST} values are concentrated at *PPARD* (Fig. 2a), with the peak (all F_{ST} values above 0.5) located in the downstream half of the gene body. The distribution of F_{ST} values displays a notable contrast between both halves. A spot with an increased recombination rate coincides with a boundary for the different sequence signatures (Fig. S7). Accordingly, we hypothesized that the coincidence of selection and recombination at *PPARD* resulted in the molecular profile around the genomic region and the genetic stratification within the populations (Fig. 2b). The original gene underwent fragmentation after a recombination event. The adaptive haplotype carrying the selectively beneficial allele upstream of the recombination site was linked to another downstream haplotype which was evolutionarily neutral. The adaptive haplotypes began to expand in population under selection, and the neutral alleles on the recombinants were concomitantly amplified as a consequence of genetic hitchhiking. The individuals in a population were genetically bifurcated into two groups depending on whether or not they were the recombinant carriers. Presumably, the CHB2 subpopulation is comprised of such carriers.

To validate this hypothesized evolutionary scenario, additional analyses were conducted for the sequence variations around the *PPARD* region in African, European, and East Asian populations. In classical neutrality tests, the entire gene was primarily scanned using Tajima's *D*. The results show constant negative values, suggesting selection, along the segment upstream of the recombination spot in CEU and CHB (Fig. S8). The combination with other tests, including Tajima's *D*, Fu and Li's *D* and *F*, Fay and Wu's *H*, and normalized *H* statistics, consistently indicates a significant positive selection in the upstream region (Table S7). The fluctuation of Tajima's *D* to positive values in short downstream fragments in CHB2 poses a contrast between the two Han Chinese subgroups (Fig. S8). The significant positive Tajima's *D* values in two fragments in CHB2 literally signify balancing selection or demographic contraction for an enrichment of intermediate frequency polymorphisms (Table S7). Actually, this explanation seems unlikely. The significant statistics for the two fragments, such as Tajima's *D* and Fay and Wu's *H*, favor a persistent signal of Darwinian positive selection in CHB1.

Long-range haplotype tests based on LD extension were applied to detect the signatures of positive selection for the target region (Fig. 2a). Selective evidence was identified in YRI (rs9470005,|iHS|=3.09707) and CEU (rs2284196,|iHS|=5.32864) with the top signals lying within the identical upstream interval. The derived alleles are selectively favored. The iHS evidence appears weak in CHB, but the cross-population extended haplotype homozygosity (XP-EHH) metrics [31] offer suggestive evidence of selection. The XP-EHH signals emerged on the both immediate sides of the iHS peak

when comparing the haplotype extension between YRI and CHB (YRI-CHB, YRI-CHB1 and YRI-CHB2). Notably, the downstream XP-EHH signals were revealed in the analyses with CHB1 (YRI-CHB1 and CHB1-CHB2). The XP-EHH results between CEU and CHB (CEU-CHB, CEU-CHB1 and CEU-CHB2) probably endorse selection with a comparable strength in both populations. The XP-EHH between the two CHB subpopulations proves a longer haplotype extension in CHB1, as evidenced by a stronger XP-EHH stretch along the downstream region.

Median-joining networks for the *PPARD* haplotypes present consistent results. The star-like networks (Fig. 2c), due to an expansion of closely related haplotypes, support positive selection on this locus, especially in CEU and CHB. In the networks for the downstream of the recombination, a haplotype cluster, Haplogroup B, is distinctively derived from another cluster, Haplogroup A, which represents an iconic pattern of positive selection in a population. Moreover, the haplotypes belonging to Haplogroup B are devoid of those from CHB1 and account for nearly half of the total CHB2 haplotypes. This equivalent is missing in the networks for the upstream. Haplogroup B justifies the recombinant carriers of CHB2 and the genetic hitchhiking effect from selective sweep. The high frequency (> 94%) of Haplogroup A in CHB1 and the combined frequencies from Haplogroup A and B in CEU and CHB, approaching 90%, inform an identical natural selection in CEU and CHB, as Haplogroup A and B represent the selectively favored haplotypes. The upstream networks imply a selection on standing variation. In addition, the median-joining networks constructed using 21,280 haplotypes from the CONVERGE data [32] with a broad coverage of the Han Chinese populations bolster the evidence of selection in East Asian population (Fig. S9). The networks for the two divided regions by the recombination certify the genetic hitchhiking effect of selective sweep from standing variation.

Despite the existing tests for selection with variable power to a range of selective scenarios, they individually remain deficient in effectiveness to resolve the current situation. A progressive combination of the analyses uncovered the contrastive molecular signatures in sequence and proved a positive selection on *PPARD*. It could be reckoned that a causal allele conferring selective advantages is situated in the upstream region. As yet, a probe into the evolutionary history would be indispensable to infer how the actual selective forces arose.

Evolutionary history of the genetic adaptation

The clarified selection footprints left in genomic sequence necessitated an understanding of the trajectory of the genetic adaption, the selection strength and its population coverage to facilitate a comprehensive functional characterization of this adaption driven by the selective driving force.

Recent advances in genealogical inference have paved a way for ancestral recombination graph (ARG) based methods used to address many fundamental questions in population genetics [33–35]. A full ARG captures highly informative features that are able to be leveraged in accurate inferences about selection, as it presents the order and time of coalescence of local genealogies that recapitulate shared evolutionary history and recombination events for a collection of sequences of interest. Selective effects acting on an allele have been allowed to characterize from ARG with agreater power, based on departures from the expected genealogical patterns of coalescence and recombination under neutrality [36–37]. Inferred local trees, represented by topology and branch length, are robust enough to test a given haplotype block under selection, even with recombination disturbance. Sweep-driving SNPs are expected to yield more significant signals and to be separated the from linked neutral alleles by recombination. We adopted a genealogy-based framework combining Relate [33] with CLUES [36] as an efficient approach for information-rich inference about selective history, which may boost the understanding by summarizing haplotype diversity over long genomic regions.

We utilized the united Han Chinese population of CHB and CHS (CHB + CHS), CEU, and YRI as model populations for inference. To secure positive selection signals robust to spurious latencies from genetic drift and demographic events, we first approximated the demographic history of each model population. Relate built up the coalescence rates over time to obtain a neutral expectation and traced the specific effective population sizes for Chromosome 6 where *PPARD* resides, in that the forthcoming analyses for our target are vulnerable to demographic dynamics. With the selection inferences adjusted for the population-resolved demographic changes in history (Fig. S10), the positive selection statistics were assessed in each model population. The stronger evidence was manifested in the range overlapping the signals in the long haplotype tests (Fig. 3a). The significance for CHB + CHS is evidently the strongest among the model populations, bolstering the suggestive evidence of XP-EHH for a comparable selection between CHB and CEU. The top hits, rs9470007 ($P = 9.52 \times 10^{-6}$) for CHB + CHS, rs3777744 ($P = 1.42 \times 10^{-4}$) for CEU, and rs11967065 ($P = 5.49 \times 10^{-4}$) for YRI, were chosen to extract their local trees of interest. rs9470007 and rs3777744 are close to and in strong LD with rs2284196 identified in the iHS test (Fig. 2a). The local trees of the focal SNPs concordantly depict much higher coalescence rates for the derived alleles in CHB + CHS and CEU (Fig. 3b). Specifically, their deformed topology and

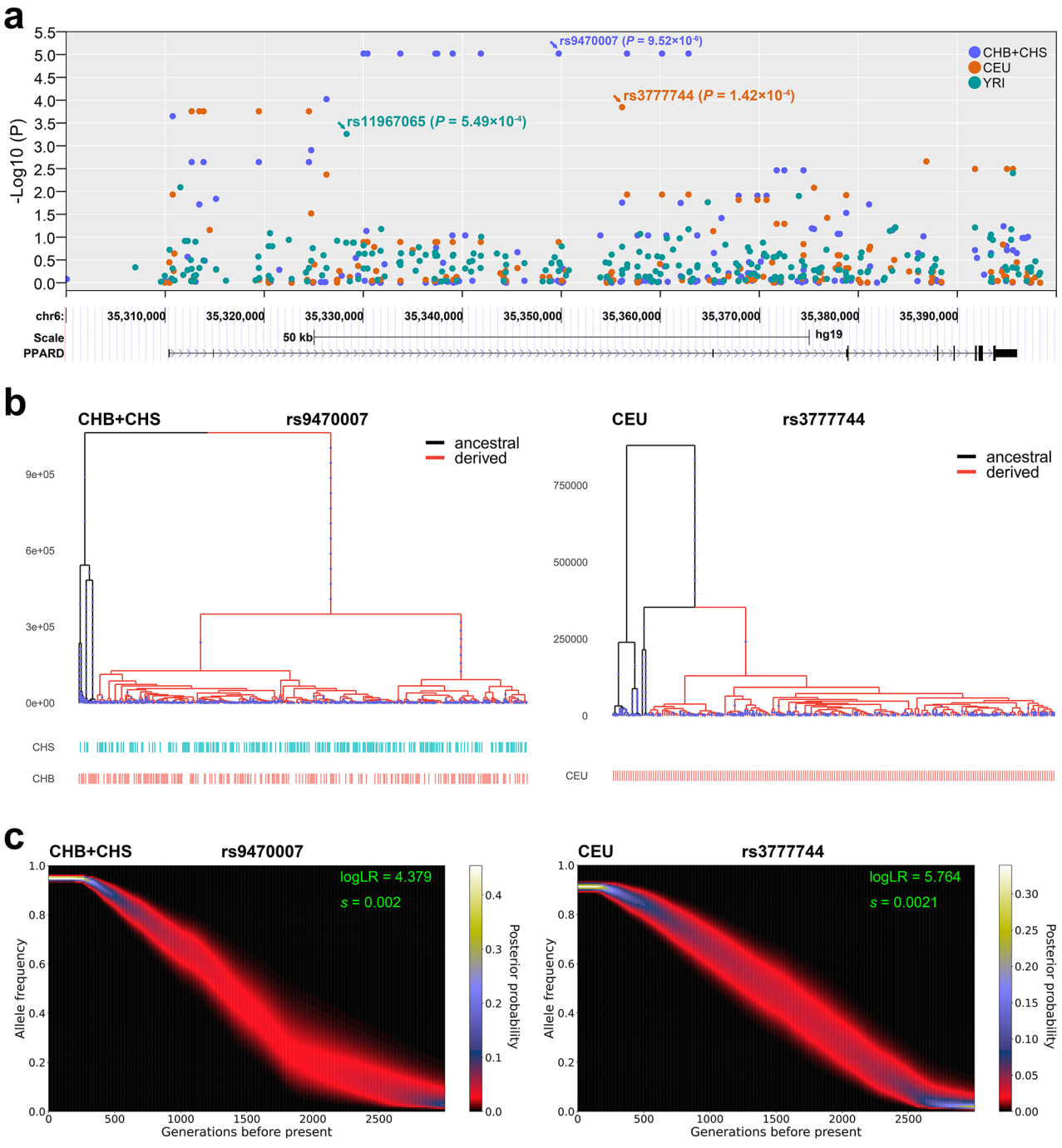


Fig. 3 Selection test and evolutionary inference of *PPARD* from a genealogy-based framework. **a** The regional plots of the selection tests in the combined Han Chinese of CHB and CHS, CEU and YRI populations using Relate. The most significant SNPs in each population are marked. **b** The extracted local trees for the SNPs of interest. The left part indicates the local tree for the top hit rs9470007 in CHB + CHS, and the right indicates the local tree for the top hit rs3777744 in CEU. The branches carrying the ancestral or derived allele are charted in different colors. **c** The approximate full-likelihood inferences of selection for the SNPs of interest using CLUES. The frequency trajectories of the selected alleles were inferred after resampling the genealogies of interest, and the log-likelihood ratios (LogLR) between selection and neutrality and the selection coefficients (*s*) were computed. The allele frequency trajectories are plotted against time in generations before the present

shorter length of branches carrying the derived allele mean deviation from neutrality expectation. rs2284196 displays an identical case as rs3777744 in CEU (Fig. S11). The coalescence from the local trees allowed us to deduce that the selective sweep acted on standing variation and drove the derived allele up to a high frequency.

In the streamlined analyses following Relate, CLUES utilized the SNP genealogy properties to approximate the full likelihood of selection and infer the selected allele frequency trajectory and selection strength. The log-likelihood ratios (LogLR) give an easy rejection against neutrality for the target SNPs in the Han Chinese population (rs9470007, LogLR = 4.38) and CEU (rs3777744, LogLR = 5.76) after resampling the genealogies of interest (Fig. 3c). The allele frequency trajectories were consistently traced in CHB + CHS and CEU. The derived alleles probably had been occurring at a low frequency before the expansion in population, and the frequencies began a stable growth 2,500 generations before present (approximately 70,000 years ago). The derived allele frequency reached as high as 0.96 for rs9470007 in CHB + CHS and 0.92 for rs3777744 in CEU. Therefore, the evidence from evolutionary history supports a positive selection from standing variation and suggests an onset time of selection in appropriate alignment with the early out-of-Africa migration of modern humans. The identical estimates of a maximum-likelihood selection coefficient, s≈0.002 for rs9470007 in CHB + CHS and s≈0.0021 for rs3777744 in CEU, prove a moderate selection strength based on a work for the classification of selective sweeps using machine learning [38]. Meanwhile, a discrepancy was observed in YRI. The local tree of rs11967065 and the CLUES inference present moderate evidence for a strong selective sweep on a *de novo* mutation (LogLR = 3.31, s≈0.041), suggesting a much more recent event (Fig. S12).

We performed further analyses to validate whether the long-lasting selection occurred in extensive Eurasian populations, in that its onset timing and the allele frequency history appropriately correspond to the peopling across Eurasia. Thereby, broader populations of Eurasian representatives were interrogated by leveraging the CONVERGE data for East Asians and all the European populations (EUR) in the 1KGP data. Given the trade-off between computational affordability and potential benefits from increasing sample size, we subsampled a collection of 1,500 random individuals (3,000 haploids) from CONVERGE for analysis. Evidence of positive selection for *PPARD* was substantiated in major Eurasian populations across continents. rs9470007 ($P = 1.87 \times 10^{-7}$) and rs3777744 ($P = 3.64 \times 10^{-4}$) remain among the top hits in the Han Chinese populations from CONVERGE (Fig. S13) and EUR (Fig. S14), respectively. A clearly elevated coalescence rate for the derived alleles in the local trees was translated to strong evidence for moderate selection in CONVERGE (logLR = 9.78, $s = 0.0024$) and EUR (logLR = 11.36, $s = 0.0023$) after the CLUES inference with a correction for individual population demography. We have been allowed to propose a common genetic adaptation across the populations outside Africa.

In consequence, we legitimately conceived that a selectively beneficial allele was shared among major Eurasian populations during their dispersal out of Africa, from the evolutionary history and coverage of the genetic adaptation. It became essential to determine where the selective driving force on the *PPARD* locus came from.

Biological functions of the selected allele for a prioritization of causality

The advantageous allele ascending to a high frequency over its evolutionary history would denote the critical roles in enhancing survival fitness of the out-of-Africa humans. The questions, including what forces drove the selective pressure, whether associated biological functions are in alignment with the evolutionary history, and whether *PPARD* is the effective gene inducing the selection, were in need to be addressed.

Although the putative adaptive haplotypes are consistent between the CEU and Han Chinese populations, the true variant responsible for this selective advantage remained undetermined. The SNPs with the strongest signal could be used as a rational proxy. To better characterize the driving forces for the genetic adaptation, we collated a wealth of complex traits by inquiring the association records out of 42,350 genome-wide association study (GWAS) datasets (May 2023) hosted in the OpenGWAS database [39] using the tag SNPs, rs9470007, rs3777744, and rs2284196, for a phenome-wide association (Phewas) analysis. The associations were retained with P value $< 10^{-3}$, and similar querying results were obtained for the tag SNPs (Fig. 4; Fig. S15-S16; Table S8-S10). The tag SNPs serve as expression quantitative trail loci (eQTLs) of a string of the nearby genes surrounding *PPARD* in whole blood. The associations are most significant with the expression of *DEF6* ($P = 4.44 \times 10^{-256}$), *BLTP3A* ($P = 3.52 \times 10^{-189}$), *ZNF76* ($P = 1.05 \times 10^{-51}$), *SNRPC* ($P = 5.71 \times 10^{-24}$), and *ILRUN* ($P = 3.24 \times 10^{-21}$). Other associated complex traits are notably involved in immunity, inflammation, metabolism, and physical capacity, which are frequent causes of natural selection. We classified these traits into several categories: (I) immunity, inflammation and hematogenesis; (II) physical capacity; (III) metabolite levels; (IV) energy and lipid metabolism; (V) protein levels in immune system and (VI) mental, socioeconomic status or behavior. The associations of a few traits pertaining to blood cells and physical capacity are the most significant, exceeding the genome-wide threshold ($P < 5 \times 10^{-8}$). The blood cell traits and the plasma protein levels have tight

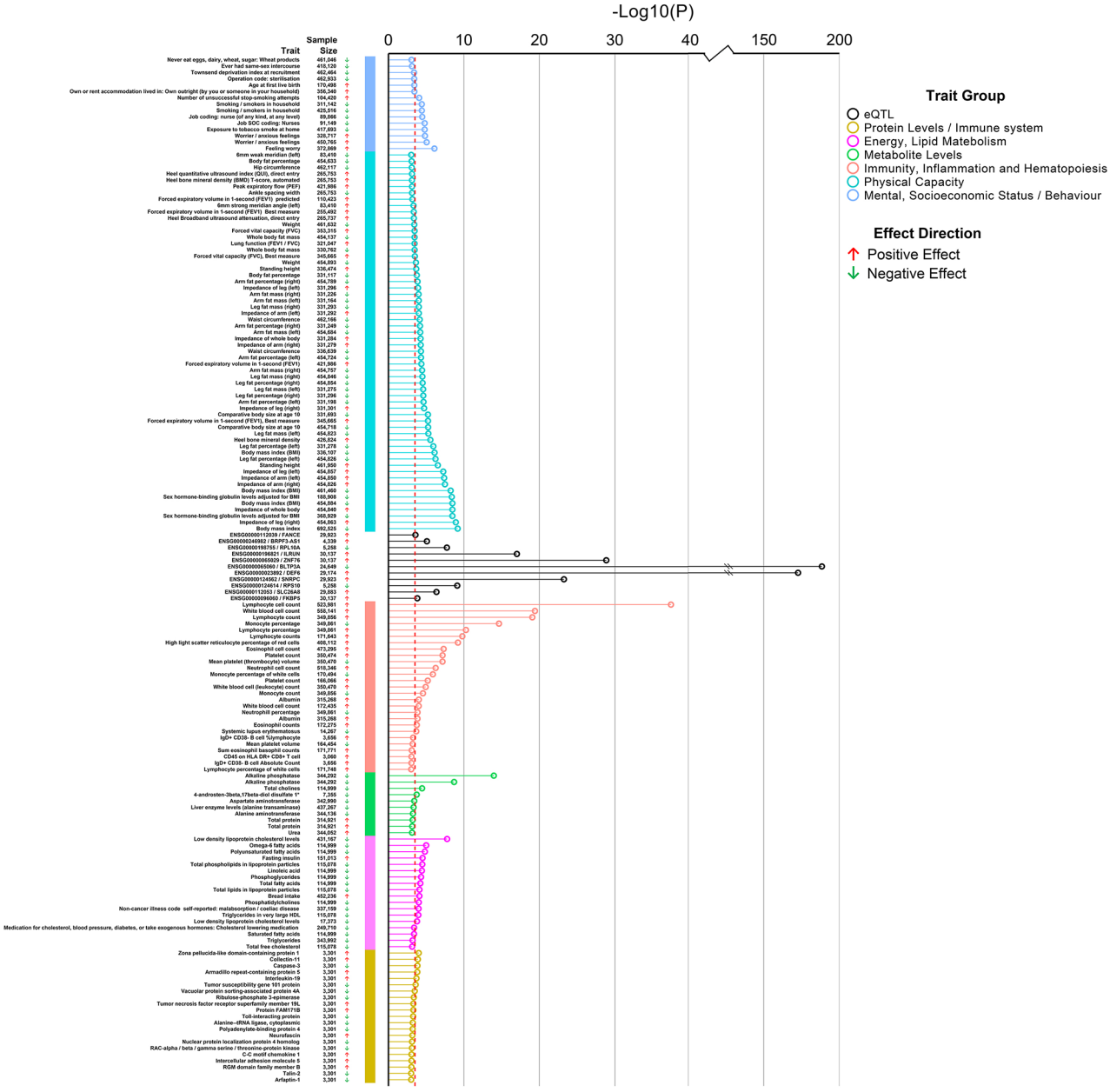


Fig. 4 Phenome-wide association analysis of rs9470007. The association records were retrieved with *P* value $< 10^{-3}$ out of 42,350 GWAS datasets in the OpenGWAS database. The associated traits and the corresponding significance are plotted. The association records of the same traits from different GWAS datasets are included. The sample size in each GWAS dataset is also listed. These traits are categorized into several groups. The dashed red line denotes a Bonferroni-adjusted significance level ($P = 3.09 \times 10^{-4}$) if corrected for 162 phenotypes

connections with immunity and inflammation. A few traits related to mental, socioeconomic status or behavior are unexpectedly observed, and they would be the consequence mediated by other exposure traits due to a supposedly less genetic contribution to such traits. Although certain phenotypes were assigned into a category, they probably interact with the traits of other categories, and some of the phenotypic outcomes could be mediated through other traits as exposures. Sex hormone-binding globulin (SHBG) levels with a strong genetic association ($P = 1.0 \times 10^{-9}$) in the Phewas results was identified to affect types of body mass, fat measures of different body parts, fasting insulin levels, muscle strength, sexual function and bone mineral density [40]. These corresponding traits are also present in the Phewas results. The a current omnigenic model has become dominant [41]. A large number of small effect size variants cumulatively contribute to the significant phenotype variance. Generally, the Phewas results reflect the pleiotropy conferring favorable effects on the associated complex traits.

The prominent eQTLs signify extended *cis*-regulation on the local gene expression. We postulated that complex traits might be mediated through the regulatory effects on gene expression. The association significance in the Phewas data could deliver an informative clue to selective pressure. Furthermore, it would be necessary to continue to explore the underlying genetic determinant. A common genetic adaptation across major Eurasian populations allowed us to envisage an independent association underlying the spectrum of phenotypes. Thus, we conducted a Bayesian colocalization analysis with coloc under a single causal variant assumption [42], seeking to compare the correlation between individual traits and gene expression, map the likely adaptive traits, evaluate the shared genetic causality, and provide evidence for the true causal variant. A statistically independent locus was defined, comprising the intersected variants in strong LD with rs9470007 ($r^2 > 0.75$) in the CHB + CHS and CEU populations. The trans-ethnic intersection yielded a parsimonious set of candidates, including 15 SNPs spanning 52 kb in the genomic region (Table S11), which maximally represent a unique significant association in the GWAS data. The summary statistics of these variants were extracted from the eQTL and GWAS datasets for the following colocalization analyses. The GWAS datasets with a larger sample size were chosen for the same traits, leaving 109 traits into the analyses with eQTL data of the putative effector genes (eGenes).

We first performed the colocalization analysis with the eQTL data of eGenes from whole blood. The PP.H4.abf values of colocalization were used to prioritize the most likely eGene-trait pairs. The colocalized results demonstrate a cluster of GWAS traits with a high posterior probability of PP.H4.abf > 0.85 (Fig. 5a; Table S12). The clustered traits with the highest probabilities are more enriched in two categories, physical capacity or immunity, inflammation, and hematogenesis. The traits of physical capacity include impedance of different body parts, body mass index, fat mass, height, heel bone mineral density, and SHBG, presenting outstanding PP.H4.abf values. The blood cell traits regarding lymphocytes, red cells, platelets, neutrophil cells, and monocytes are prominent from the category of immunity, inflammation, and hematogenesis. A common *cis*-regulatory role of the putative selected variant in the expression of these eGenes might leave the analysis insufficient to discriminate the correlation between the exact eGenes and the mediated traits. However, it allowed prioritization of the most likely causal variants (Fig. 5b; Table S13). Two SNPs, rs6457816 and rs73413718, yield more significant posterior probabilities after the colocalization analyses for all traits. The PP.H4 values approaching 1 for rs6457816 and rs73413718 were observed across all traits through the mediating effect of *BLTP3A* and *DEF6*, respectively. Two genes hold the strongest eQTL significance.

Although the large sample sizes rendered a high statistic power in the analyses with the eQTL data of whole blood, the single tissue source data failed to evaluate the mediating effects of gene expression on traits across diverse tissues. We further introduced the eQTL data of more various tissues from the EBI eQTL catalogue [43], a resource of quality-controlled, uniformly processed gene expression and splicing QTLs. The datasets for across tissue colocalization analysis were compiled using two criteria for statistical power consideration: (I) tissues with a sample size > 300 (Table S14); (II) at least one eQTL *P* value $< 10^{-3}$ for the candidate SNPs (Fig. S17). Forty-two eQTL datasets, comprising of eGenes from a broad range of tissue sources, were retained. The colocalized lay-out between the eGenes across tissues and the GWAS traits maintained similar prioritized traits with high probabilities of PP.H4.abf (Fig. S18; Table S15). The traits belonging to the two groups, physical capacity and immunity, inflammation, and hematogenesis, are largely reproduced with the highest probabilities. Two eGenes, *BLTP3A* and *DEF6*, are prioritized as most likely in the colocalization across many tissues, which is attributable to their higher eQTL significance. Unexpectedly, the *PPARD* gene is significantly colocalized in fibroblasts, which make interpretation for the relevance to the GWAS traits elusive. The two SNPs, rs6457816 and rs73413718, are still prioritized with higher PP.H4 values across the major tissue-eGene pairs (Fig. S19; Table S16). Despite lack of the exact mechanism by which eGene contributes to a colocalized trait, the above analyses illustrate that the putative selected variant would act as a regulatory hub for gene expression in a local genomic region, and shed light on the pleiotropy which may lead to multiplex selective forces driving the genetic adaptation.

Discussion

We detected an exceptional sub-population differentiation arising from alleles at the *PPARD* locus after the initial characterization of the genetic architecture in the ADME genes among the representative Chinese populations (Fig. 1; Fig. S5-S6). A group of functionally specialized genes led to the identification of an unexpected genetic distinction, in contrast to the general population structure obtained from unbiased genetic markers [24]. It was further verified as an authentic result, presumably owing to natural selection, using the 1KGP data of the intercontinental populations. We applied progressive assays combining extended haplotype homozygosity, phylogenetic haplotype networks, and frequency spectrum-based statistics to illuminate the sequence signatures shaped by selective sweep (Fig. 2; Fig. S8-S9; Table S7). The signatures manifest a sharp contrast across the

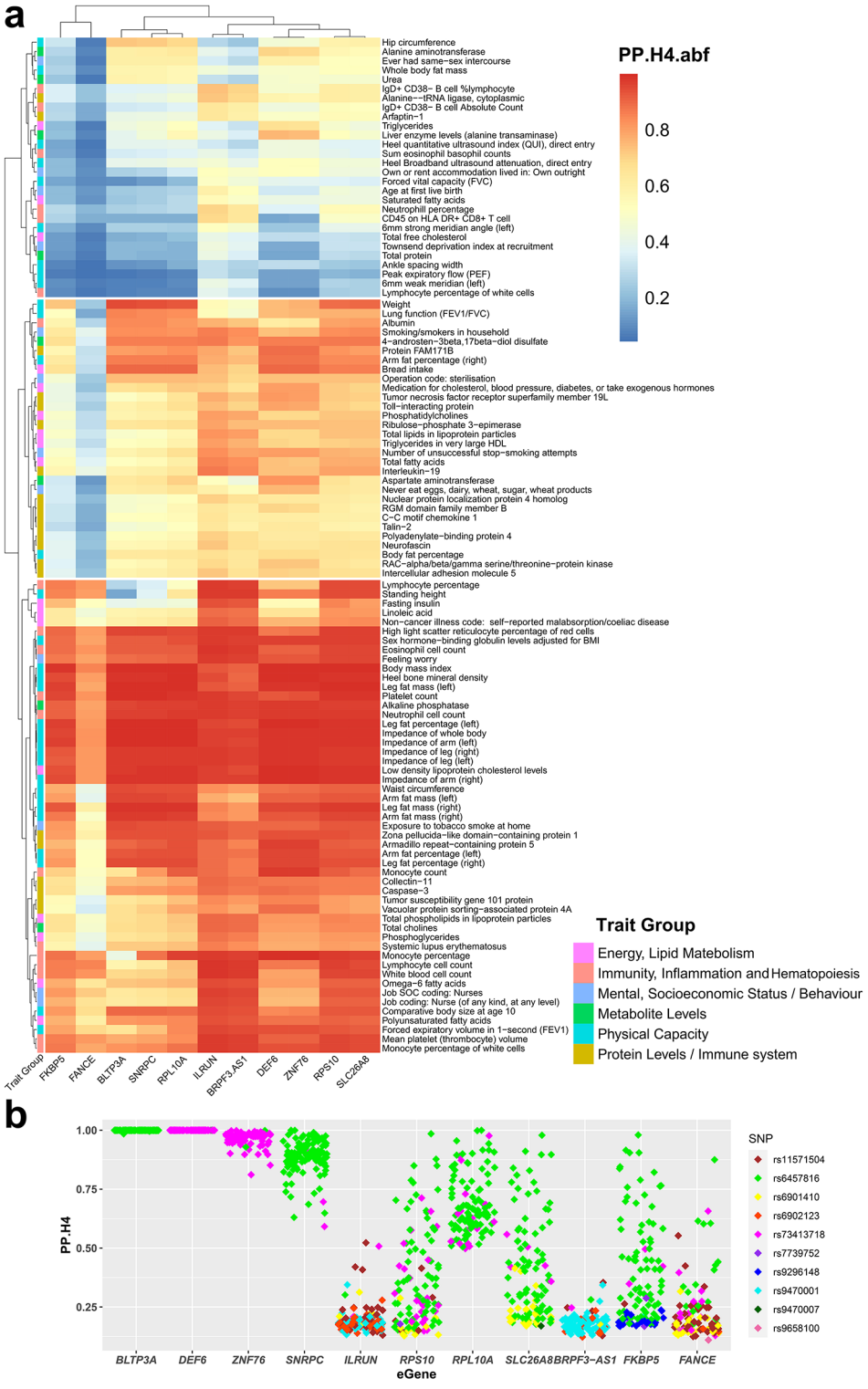


Fig. 5 Colocalization analysis between 109 GWAS traits and the eQTLs from whole blood. The prior parameters $p1 = 10^{-4}$, $p2 = 10^{-4}$, $p12 = 5 \times 10^{-6}$ were adopted in the coloc.abf function for colocalization. **a** The clustered heatmap of the colocalized posterior probabilities of PP.H4.abf between the traits and the eQTLs. Heatmap colors correspond to PP.H4.abf values. The GWAS traits are arranged in rows, and the eGenes are in columns. The trait categorization is identical to that in the Phewas analyses. **b** The posterior probabilities of the most likely causal variants from colocalizing between each eGene and the GWAS traits. Each diamond stands for a variant with the highest posterior probability in each colocalization analysis. The prioritized variants are denoted in different colors

partition of a recombination spot. Recombination usually disturbs the robustness of such analyses when examining positive selection. Thus, efficiency from each of the tests alone could be diminished. Besides, this class of approaches through scanning the patterns of sequence variation for extreme outliers are limited in their ability to eliminate a chance induced by demographic transition. However, by comparing the separate statistics from both sides of the recombination spot, our combined results strongly approve a sign of selective sweep on this locus, because they effectively overcome the confounding effects from demographic processes that operate at a genomic scale rather than on segregated stretches in a local region. The molecular signatures from the upstream and downstream sequences of the recombination spot alleviate the confounding concern, as the paradoxical profiles instead highlight a credible genetic hitchhiking effect of "linked selection", and discriminate it against the resemblance mimicked by background selection that also creates skews in allele frequencies at linked sites nearby. Moreover, a genealogy-based framework figured out the evolutionary history from the genealogical inference of *PPARD*, indicating that a positive selection had been operating across the Eurasian populations since early modern humans began a long journey outside Africa and that a beneficial allele has ascended to a high frequency close to fixation (Fig. 3; Fig. S13-S14). The selection onset timing could pose another doubt regarding whether the high frequency of the derived alleles resulted from bottleneck effect [44]. The correction of population-resolved demographic dynamics from the estimated effective population sizes (Fig. S10) in the inference over selective history has eliminated this problem. The signals targeted by various selection scans may overlap in the immediate vicinity of a genomic stretch across multiple populations. We need to distinguish these selective events caused by either convergent adaptation or common ancestry potentially under common selective pressure. The consistent patterns of the sweeping haplotypes and coalescent traces across the Eurasian meta-populations should be more concordant with the latter. Meanwhile, a discrepancy observed in African YRI (Fig. S12) is justifiable. Adaptation from standing variation may be less distinguishable from the background noise of genetic drift, especially in local populations. The constant increase in the derived allele frequencies over times actually reflects adaptive fitness among extensive populations and controls a spurious risk due to genetic drift in population history.

A few disadvantages might leave the sophisticated scenario of *PPARD* underappreciated before. Gene flow has been distinguished as the primary contributor to human allele frequency changes over the last few thousand years, with genetic drift but not selection accounting for much of the remaining part [45]. Admixture events, which prevail in complex human population history, can obscure sweep signals and perturb accurate inference about selection [46]. Canonical models of hard sweeps postulate that a de novo beneficial mutation emerges under a ready driving force, marking the onset of positive selection. Higher power seems available to detect ongoing partial sweeps from recent selection [47], in which selection quickly drives an advantageous allele along with the sweeping haplotypes to higher frequency, generating obvious deviations in variation patterns from the rest of the genome. Thereby, they can be easily discerned by the popular algorithmic models. The rapid LD decay of haplotype extension over time would instead undermine a sweep signal. Recent theoretical and simulation-based studies have shown that selection may trigger the divergence of standing genetic variants during environmental shifts [48], signifying a change in chronological order for the emergence of genetic variation and selection pressure. Standing genetic variation under selection enable a better adaptive response to driving force, such that this selection model, namely soft sweep, may play a substantial role in adaptation to new environments during human evolution [49–50]. Yet, this fundamental evolutionary force is deficient in convincing evidence in practice, largely because the tools, that are ubiquitously efficient in tackling the sophisticated scenarios of soft sweep, remain lacking, despite theoretical and data-driven advances over the past few years [51]. The integrative application of the diverse analyses, which are suited to various selective scenarios and are more adept at tracing certain aspects about selection, added to the required evidence. From the genomic datasets of the contemporary populations, we endeavored to achieve an unbiased understanding of the selection mode and tempo, which may imply the intriguing case of *PPARD* seems intermediate between hard and soft sweep. Compared to strongly selected alleles of young age in hard sweep, ancient ones are easier to approach fixation but more likely to miss in most analyses owing to less power [46]. In the presence of the signals from haplotype signatures, the evolutionary trajectory from coalescence allowed us to unravel a distinctive episode of positive selection in the present-day humans across Eurasia. Additionally, the results are robust to challenges from the major migration and admixture events of ancestral populations, which compounded genomic variations after ancestry transitions in history.

Modern humans encountered fresh and diverse environments when they migrated out of Africa and began to disperse across alien Eurasia. The selections that have been cumulatively observed for more adaptive traits tend to be typical hard sweeps [47, 52]. Less salient signatures in the human genome, which might be left by the dominant modes of adaptation, such as soft sweeps and

polygenic selection, have not been well characterized [50]. We distinguished the persistent sweep signal that could endure historical demographic dynamics, such as bottlenecks, expansions, and admixtures [46]. The sweep most possibly originated from early out-of-Africa ancestors before the subsequent Eurasian separation, and was maintained in the vast majority of Eurasian populations rather than in certain localized cohorts, suggesting critical roles that the selected locus played in the biological processes for better adaptation, and sustained selection pressures over a long evolutionary history. Ancient DNA has become an indispensable complement to modern genomic data, which is insufficient for establishing past selection events, although the defects in data integrity are frequently inevitable. A recent study exploring global ancient human genomes originally reported *PPARD* as one of the high-confidence sweep signals, and proposed a termed "Arabian Standstill" model that the ancestral out-of-Africa group underwent a prolonged period of genetic isolation and adaptation outside Africa, presumably in the Arabian Peninsula area, potentially from at least 80,000 years ago until the subsequent Eurasian expansion 50,000 years ago [53]. Our dating inference for this gene exactly aligns with the ancient genomic evidence. The uniform evolutionary trajectories that we reconstructed from the modern descendants of separate Eurasian populations approve the selection onset of *PPARD* fall within the Arabian Standstill phase. Therefore, this model reinforces the interpretation to our results because of a better understanding about the evolutionary history of the locus and why the selection occurred in transcontinental populations. Our findings, in turn, illustrate evidence of an exemplified gene locus through coalescent inferences from the genomic data of contemporary populations to this model.

Although positive selection signs have been mounting from studies in diverse populations worldwide, most of them remain agnostic to the underpinning selective pressures. It is challenging to systematically understand genetic adaptation, unless persuasive evidence has correlated a functional variant from the putative adaptive haplotype with precise biological processes and phenotypes under the corresponding selective forces. To this end, we incorporated the Phewas datasets (Fig. 4; Fig. S15-S16; Table S8-S10) with the eQTLs (Fig. S17; Table S14) across a diverse set of body tissue/cell types to better characterize the biological mechanisms underlying organismal adaptation. The genetic target under selection should represent a regulatory element located at *PPARD*, responsible for a hub in the regulation of gene expression via *cis*-acting effects on the nearby genes that mediate influences on the downstream phenotypes through individual pathways. The colocalization results (Fig. 5; Fig. S18; Table S12-S13; Table S15-S16), albeit less explicit and stringent, could intensify the validity of the selective forces elicited from physical capacity, energy metabolism, and immune-related involvement. The over-represented adaptation traits can be attributed to the pleiotropy of the putatively selected variant. Notably, PPAR-δ alone plays pluripotent roles with a broad expression spectrum including liver, skeletal and heart muscles, skin, gut, placenta, adipose tissue, and brain [54]. It participates in the control of carbohydrate and lipid metabolism, reduces lipogenesis, alleviates inflammation and endoplasmic reticulum stress, ameliorates insulin resistance, and lowers the expression of inflammatory cytokines. Globally, *PPARD* can govern crosstalk between the metabolic and innate immune systems [27]. Its essential roles may enable a proper explanation to the overrepresentation of adaptive traits. Although the *cis*-regulation of the *PPARD* expression appears less significant across tissues and cell types in the available eQTLs, its low abundance with minor variability would be sufficient to execute the downstream functions as a crucial transcriptional sensor. Moreover, the shared genetic determinant evidently orchestrates the expression of a few genes in a local region, among which *BLTP3A* and *DEF6* are the most significant. Both genes are extensively expressed, but their biological roles remain poorly characterized. The genetic associations of *BLTP3A* were identified with systemic lupus erythematosus risk [55–56] and fasting insulin levels adjusted for body mass index [57] at a genome-wide significance. The risk-raising allele of systemic lupus erythematosus was also positively correlated with the gene expression [56]. *DEF6* participates in the modulation of adaptive and innate immunity, with particularly pivotal roles in regulating development, activation, and function of T lymphocytes [58–60]. Def6-deficiency resulted in inflammation, autoimmune disorder symptoms, and diverse immune defects including abnormalities in T-cell expansion and T-helper cell differentiation, sever hypergammaglobulinemia, and autoantibody production in mice [61]. The missense mutations in *DEF6* brought about the carriers with immunodeficiency and systemic autoimmunity because of the *DEF6* role in regulating abundance and recycling of the T-cell checkpoint protein CTLA-4 [62], which is critical for human immune homeostasis [63–64]. A strong genetic signal in *DEF6* was also found associated with the onset of systemic lupus erythematosus [65]. The associated variant is the *cis*-eQTL of *ZNF76* and *DEF6* and the ancestral allele is hazardous to the disease. High *DEF6* expression levels proved a potential biomarker for tumorigenesis, metastasis, and poor prognosis in multiple cancers [66]. Expression of the local genes is prominently colocalized, even across multiple cell and tissue types, with a group of traits that may be adaptive to better fitness in the present analyses. Noticeably, the positively selected derived

allele renders pleiotropic effects on the associated complex traits almost in a favorable direction by boosting benefits and inhibiting drawbacks (Fig. 4; Fig. S15-S16; Table S8-S10) to attain a harmonized fitness. Accordingly, the estimated selection coefficients may represent a composite pressure resulted from the interplay of various driving forces during the past evolution. A natural request in future research would be to precisely partition the pleiotropic effects from individual mediating eGenes responsible for the corresponding traits, and to ascertain the adaptation mechanisms.

Clarifying the biological mechanisms by which target genes in critical tissues give rise to a phenotypic outcome remains a challenge, particularly when complicated by pleiotropic effects from a locus that can independently impact gene expression and traits. Co-regulation of genetic effects on gene expression across tissues impedes the disentangling of truely causal effects from other concomitant effects, as shared causal eQTLs or LD with causal eQTLs typically fail to differentiate co-regulated genes and tissues in analyses. The latest methods have been even able to jointly model genetic co-regulation across tissues for heritability enrichment, partition individual effects of gene expression on complex traits from co-regulated genes and tissues, and statistically distinguish causality from proxies [67–68]. In practice, they may deliver more benefits in the presence of gene expression in a tissue- or cell type-specific manner [69–70]. The eQTL data has informed that the *cis*-regulated eGenes span a genomic range exceeding a reach within which regular fine-mapping methods are more adept given multiple causal genetic elements at a target locus. Then, we conducted the colocalization analyses with simplified requirements, such as an LD matrix reference or even in-sample genotype data. An independent, LD-defined candidate set (Table S11) was speculated to harbor a unique causal effect of the putative selected variant. The coloc algorithm calculated the posterior probability of a shared causal variant between a GWAS trait and expression variation from a single tissue-gene pair. We integrated the posterior probabilities (Table S12-S13; Table S15-S16) and prioritized the putative genetic determinants (Fig. 5; Fig. S19) throughout the tissue-gene contexts with delicate allelic heterogeneity. Our practice did not follow a recommended workflow for coloc analysis, but could resemble a manner inverse to the masking method [71]. It clumped a candidate variant set in compliance with the single causal effect from the selective signal into analysis. It was presumed to preclude other distinct causations in colocalizing and to maximally narrow down the candidates for the genetic determinant. Joint analysis across association statistics is beneficial for discriminating the non-coding regulatory variant underlying complex traits, because it promotes the identification of the causal variant for GWAS signals shared in multiple phenotypes. Although the colocalization analyses have not revealed the cognate genetic causation underlying each trait, the lines of evidence above have convergently established an independent causal effect underpinning the genetic associations. The prioritized variants and effective genes with the most likely posterior probabilities were obtained from the colocalization analyses. The high probabilities should be conservative evidence for the colocalized links between gene expression regulation and GWAS traits, and robust to refine the causal variant.

Moreover, appreciation for the divisive signatures in the sequence left by the selective sweep and the recombination offered valuable information for fine-mapping the driving variant. The characteristics for such variant, presumably localized upstream of the recombination, with a high frequency of the derived allele, and in strong LD with the surrogate SNPs, would facilitate the recognition of the causal variant. High frequencies of rs9470007 derived allele among major Eurasian populations (Fig. S20) suggest that it is more analogous to the genuine variant and suitable for clumping the LD-defined candidate set. The both refined variants, rs6457816 and rs73413718, precisely conform to the presumed properties of the genuine target deduced above.

The mediating mechanisms of causal variants and genes usually keep silent and undetected because of the lack of context-specific effects for genetic associations [72]. The eQTL data, deficient in full coverage of cell and tissue sources, were not sufficient to comprehensively acquire the mediating effects of expression on complex traits. The unavailability of critical tissues may be another limitation for our colocalization analysis, such as the expression data in bone marrow responsible for the hematologic traits. This *PPARD* locus presents a case of pleiotropy that is more sophisticated than expected. Therefore, partially assessing the mediating effects of gene expression on complex traits is a visible restriction in our analyses. However, the prioritized results should be robust from the bulk tissues comprising miscellaneous cell types. Our study has laid a solid foundation for following efforts to elucidate the regulatory mechanisms underlying gene expression and the mediating effects on biological functions, and to make full sense of a gene-tissue-trait-adaptation picture.

Conclusions

This study highlights an overlooked episode of genetic adaptation at the *PPARD* locus, which plays a crucial role in human survival fitness. While positive selection at this locus has been well documented, the underlying biological mechanisms linking genetic variation to gene expression, complex traits, and disease susceptibility remain to be fully elucidated. A deeper understanding

of these mechanisms will be essential for advancing our knowledge of how the selectively advantageous allele contributes to adaptation. Moreover, the inferred selective history of this locus provides insights into the dynamic interplay of multiple evolutionary forces. Identifying causal variants and effective genes will be a key step toward bridging current knowledge gaps and further exploring the pleiotropic role of this locus in various adaptive processes via experimental validation.

Materials and methods

Population samples

A total of 1092 unrelated DNA samples from 9 representative populations in China were anonymously collected for genotyping, with consideration of ethnicity, culture, and regional distribution. The three Han Chinese populations, the largest part of the samples, were recruited from Guangdong province, Shanghai city, and Shandong province, representing the southern, central, and northern Han Chinese, respectively. The three ethnic minority populations from South China are the Lic, Zhuang, and Miao populations. The Tibetan, Mongolian, and Uygur populations, accounting for the largest minority populations in the northwest of China, were included in this study. Population membership was ascertained by the grandparent ethnicity of the subjects. The sample information is listed in Table S1.

Genotyping, data filtering, and public data compilation

DNA samples were genotyped using the AffyMatrix DMET Plus chip. Genotyping quality control (QC) required a call rate higher than the 98% threshold for each subject. We implemented a stepwise genotype QC procedure (Fig. S21). Finally, a post-QC set of 1802 autosomal SNPs were harvested for subsequent analyses. In addition, the phased genotypic data from the 1KGP datasets were introduced in our analyses [25]. The data of the large Chinese specific panel CONVERGE [32] were downloaded from the European Variation Archive with PRJNA289433.

Data analyses of population genetics

The identity-by-state pairwise distances of individuals were calculated using PLINK v1.07 software [73] for MDS analyses. STRUCTURE analyses [74–75] were carried out to decompose a range of ancestry clusters from K = 2 to 8 in the proportions of individuals and the average population proportions, with 10,000 burn-ins, 10,000 Markov Chain Monte Carlo (MCMC) iterations, and 5,000 admixture burn-ins under the Admixture, Admixture plus LOCPRIOR, and Linkage models, respectively. To obtain more reliable results, we replicated 10 times with random seeds for each STRUCTURE modeling at the same K, and then applied CLUMPP v1.1.2 [76] using its GREEDY algorithm, with 5,000 repetitions to obtain the optimal replicate alignments. The results were plotted for visualization using DISTRUCT software [77]. A stepwise AMOVA analysis based on various grouping hypotheses was conducted using Arlequin suite v3.5 [78] to evaluate the genetic diversity among groups, among populations within groups, and within populations. *P* values in each run were produced by significance tests of 10,000 permutations.

TreeMix v1.13 software [79] was used to model a maximum likelihood tree for the populations by inferring the patterns of population splits, gene flows, and genetic drift in history. The Plink package was used to calculate allele counts of SNPs for the required input data. Uygur was specified at the root position. For a more reliable tree topology, we modeled a range of inferences by setting the -k flag spanning 15, 20, 30, 50, 75, 100, 150, 200, 300, and 500 to account for non-independent nearby SNPs with LD, and by setting the -m flag from 1 up to 15 to allow for migration events. Changes in the log likelihood statistics were evaluated for the different modeling runs. The optimal number of migration edges was estimated by the Evanno, linear and SiZer methods using the companion OptM package [80] in R. The topology of the population divisions was then plotted. F_{ST} statistics per SNP were calculated to measure the genetic differentiation between each pair of populations by the most widely used Weir and Cockerham estimator [81], which accounts for differences in sample sizes between populations. The 99.9th, 99th, 98th and 95th percentiles of F_{ST} value distribution were calculated and plotted in a clustered heatmap using the pheatmap package in R.

Tests for natural selection

The phased haplotype data were retrieved from the 1KGP datasets. We conventionally chose YRI, CEU and CHB to represent African, European, and East Asian populations from the three main continents. Classical neutrality tests with different statistical preference were conducted. Tajima's *D* [82] was measured along the entire *PPARD* gene using a sliding window approach to characterize the deviation of allele frequency spectra of SNPs from neutral equilibrium. The tests for Fu and Li's *D* and *F* [83], Fay and Wu's *H* [84], and normalized *H* statistics [85] were performed with an outgroup of ancestral information. The significance of individual statistics was yielded through comparing the observed values to distributions generated from 10,000 coalescent simulations, conditional on the observed sample size and Theta (θ) values [86], assuming a standard neutral model with no recombination, using DnaSP v5.10 [87]. Surveyed values falling into the one-sided 5% tail of the null distribution were considered statistically significant.

We applied two complementary extended haplotype homozygosity based long-range haplotype tests to detect the signatures of recent positive selection. iHS [29] and XP-EHH [31] probabilistically model the likelihood of the haplotype decay due to the recombination background in human genome, and then quantify evidence of recent positive selection by the presence of long-range haplotypes. Both metrics were calculated with the selscan software [88]. The genetic distances (recombination rates) for SNPs used in the analyses were interpolated from the estimates in the HapMap Phase II data release [89], which combine the recombination maps for African, European, and East Asian populations, and are less likely to be affected by population-specific selective sweeps and demography [90–91]. The ancestral state for each *Homo sapiens* variant site was retrieved from the annotations in 1KGP datasets or determined based on the GRCh37_e71 ancestral genome from the Ensembl database release. Raw iHS and XP-EHH statistics were normalized using the norm tool packed in the software. The cut-off of both statistics <-2 or >2 was considered evidence of selection, largely corresponding to the top 5% hits from a genome-wide distribution. Positive and negative scores mean the longer haplotypes with the derived or ancestral allele background.

The haplotype networks were constructed using the median-joining algorithm by Network v4.6.1.4 program [92]. Ancestral allele state was examined as described above, and a haplotype composed of ancestral alleles for each site was used as the root.

Genealogy-based inferences

We executed the genealogy building and the coalescence rate/branch estimation procedures using Relate v1.1.9 [33]. We adopted a per-generation mutation rate of 1.25×10^{-8}, 28 years per generation, and the effective population size of 20,000, 30,000, and 50,000 haploids for East Asians, Europeans, and Africans, respectively. The human ancestor sequence of GRCh37 and the genetic map from HapMap were used as described in the selscan above. The genomic mask file was obtained from the 1KGP release. The phased genotypic data were processed as input files by the PrepareInputFiles module. Relate was used to infer the population genealogy with "--mode ALL". The historical effective population sizes were estimated by the EstimatePopulationSize module. A recommended tree dropping threshold of 0 was set, such that the branch lengths in all trees are updated for detecting positive selection. The.coal file outputs for coalescence rates were used in the ensuing analyses. Positive selection evidence from genealogy was detected by an addon-on module DetectSelection with calibrated branch lengths for the estimated population size history. The trees of interest were plotted by the TreeViewMutation module with a focal SNP specified. We implemented the coalescence rate through time estimations separately for the Han Chinese populations, CEU, all the European populations in 1KGP and CONVERGE. CHB and CHS were combined into analysis (CHB + CHS). EUR incorporated CEU, GBR, FIN, IBS and TSI. We subsampled 1,500 subjects from CONVERGE for an unaffordable computational burden with more samples in constructing genealogy and coalescence.

The CLUES [36] algorithm was used to assess the goodness-of-fit between natural selection and neutrality scenarios for the tested variant via a likelihood ratio test, and to infer the maximum-likelihood selection coefficient and the allele frequency trajectory from a local tree topology and branch lengths. The branch lengths of the local genealogies at the SNP of interest were resampled using the SampleBranchLengths module in Relate, with 200 MCMC-like iterations for CHB + CHS and CEU populations and 500 MCMC iterations for CONVERGE and EUR samples. CLUES modulated the MCMC samples with the importance sampling options: "--burnin 100 --thin 5" for CHB + CHS and CEU populations, and "--burnin 200 --thin 5" for CONVERGE and EUR samples. Namely, the preceding resampled trees were abandoned as burn-in, and the remaining trees were pruned by keeping every 5th tree to include into the importance sampling estimate. A posterior allele frequency trajectory for the focal SNP was estimated and plotted with a specified time cutoff of 3,000 generations before the present.

Phewas and colocalization analyses

Three tag SNPs, rs9470007, rs3777744 and rs2284196, were used in the genetic association queries from the OpenGWAS database [39] for the Phewas analyses. Data were filtered with a threshold of P value $< 10^{-3}$. The extended eQTL data of more various tissues were retrieved from the EBI eQTL catalogue [43]. We ran coloc v5.2.2 for the Approximate Bayes Factor colocalization analysis on summary statistics [42], with association P values, minor allele frequencies, sample sizes and case proportions (for binary traits) summarized in individual datasets as arguments. The input data for the LD-defined set of the candidate variants were extracted from each eQTL and GWAS dataset and primed for colocalization analyses. We ensured that the effective alleles and the effective allele frequencies were harmonized between datasets. A colocalization analysis was performed between the eQTLs of each *cis*-regulated gene and the GWAS statistics of other traits using the coloc.abf function under a single causal variant assumption. The recommended prior parameters [71] $p1 = 10^{-4}$, $p2 = 10^{-4}$, $p12 = 5 \times 10^{-6}$ were adopted in our practice for a trade-off between sensitivity and conservatism. From the outputs, a posterior probability PP.H4 indicates the strength of

evidence for a shared causal variant and positive colocalization between pairs of association statistics.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-025-11620-y.

Supplementary Material 1

Supplementary Material 2

Supplementary Material 3

Supplementary Material 4

Supplementary Material 5

Supplementary Material 6

Supplementary Material 7

Acknowledgements
We gratefully acknowledge all the contributors that made this research possible, all the sample donors for this study, and all the staff for their assistance in recruiting participants to the study.

Author contributions
W.S., T.Z., and J.S. jointly conceived and supervised this work. W.H. secured the study sample. C.Z., Y.W., and H.W prepared experimental materials and generated raw data. W.S. and J.S. designed the analysis framework. W.S. performed the data analysis. W.S. and J.S. wrote the initial draft of the manuscript. L.G., L.L., and W.H. provided valuable feedbacks and contributed to the writing. All authors approved the final manuscript and agreed with their own contributions.

Funding
This work was supported by the grant from Natural Science Foundation of Shanghai Municipality [Grant No. 21ZR1446100].

Data availability
We submitted the primary data or summary statistics that underlie this research in the supplementary materials. Any additional information required to reanalyze the data reported in this paper can be available upon request through communication with the corresponding authors. The phased genotypic data of the 1000 Genomes Project datasets [25] are available from the data portal of The International Genome Sample Resource (https://www.internationalgenome.org). The download link of the CONVERGE data [32] is http://ftp.ebi.ac.uk/pub/databases/eva/PRJNA289433/. The genetic recombination map is available from the HapMap Phase II data [89] release (http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/working/20110106_recombination_hotspots/). The GRCh37_e71 ancestral genome is available from Ensembl database release (ftp://ftp.ensembl.org/pub/release-75/fasta/ancestral_alleles/homo_sapiens_ancestor_GRCh37_e71.tar.bz2). The genomic mask file is available at http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/accessible_genome_masks/StrictMask/. The GWAS data in the OpenGWAS database [39] are available from its portal (https://gwas.mrcieu.ac.uk/datasets/). The eQTL data from the EBI eQTL catalogue [43] are available at its FTP download site (http://ftp.ebi.ac.uk/pub/databases/spot/eQTL/).

Declarations

Ethics approval and consent to participate
All procedures performed in this study involving human participants were approved by the Ethics Review Committee of the Chinese National Human Genome Center at Shanghai in accordance with the Declaration of Helsinki Principles and comparable ethical standards. Written informed consents were obtained from all participants before sample collection. Their personal identifiers and sample information were transformed to be anonymous in scientific research.

Consent for publication
The participants agreed with the anonymized publication of the population results.

Competing interests
The authors declare no competing interests.

Author details
[1]Yunnan Key Laboratory of Children's Major Disease Research, Yunnan Institute of Pediatrics, Kunming Children's Hospital, 288 Qianxing Road, Kunming, Yunnan 650228, P.R. China
[2]Shanghai-MOST Key Laboratory of Health and Disease Genomics, Shanghai Institute for Biomedical and Pharmaceutical Technologies (SIBPT), 2140 Xietu Road, Shanghai 200032, P.R. China

Received: 8 March 2025 / Accepted: 21 April 2025
Published online: 30 April 2025

References

1. Chen J, Zheng H, Bei JX, Sun L, Jia WH, Li T, et al. Genetic structure of the Han Chinese population revealed by genome-wide SNP variation. Am J Hum Genet. 2009;85:775–85.
2. Xu S, Yin X, Li S, Jin W, Lou H, Yang L, et al. Genomic dissection of population substructure of Han Chinese and its implication in association studies. Am J Hum Genet. 2009;85:762–74.
3. Cao Y, Li L, Xu M, Feng Z, Sun X, Lu J, et al. The ChinaMAP analytics of deep whole genome sequences in 10,588 individuals. Cell Res. 2020;30:717–31.
4. Cong PK, Bai WY, Li JC, Yang MY, Khederzadeh S, Gai SR, et al. Genomic analyses of 10,376 individuals in the Westlake biobank for Chinese (WBBC) pilot project. Nat Commun. 2022;13:2939.
5. Luo H, Zhang P, Zhang W, Zheng Y, Hao D, Shi Y, et al. Recent positive selection signatures reveal phenotypic evolution in the Han Chinese population. Sci Bull (Beijing). 2023;68:2391–404.
6. Daly TM, Dumaual CM, Miao X, Farmen MW, Njau RK, Fu DJ, et al. Multiplex assay for comprehensive genotyping of genes involved in drug metabolism, excretion, and transport. Clin Chem. 2007;53:1222–30.
7. Grossman I. Routine Pharmacogenetic testing in clinical practice: dream or reality? Pharmacogenomics. 2007;8:1449–59.
8. Welch RA, Lazaruk K, Haque KA, Hyland F, Xiao N, Wronka L, et al. Validation of the performance of a comprehensive genotyping assay panel of single nucleotide polymorphisms in drug metabolism enzyme genes. Hum Mutat. 2008;29:750–56.
9. Ramos E, Doumatey A, Elkahloun AG, Shriner D, Huang H, Chen G, et al. Pharmacogenomics, ancestry and clinical decision making for global populations. Pharmacogenomics J. 2014;14:217–22.
10. Kimmel SE, French B, Kasner SE, Johnson JA, Anderson JL, Gage BF, et al. A Pharmacogenetic versus a clinical algorithm for warfarin dosing. N Engl J Med. 2013;369:2283–93.
11. Pirmohamed M, Burnside G, Eriksson N, Jorgensen AL, Toh CH, Nicholson T, et al. A randomized trial of genotype-guided dosing of warfarin. N Engl J Med. 2013;369:2294–303.
12. Li J, Zhang L, Zhou H, Stoneking M, Tang K. Global patterns of genetic diversity and signals of natural selection for human ADME genes. Hum Mol Genet. 2011;20:528–40.
13. Bai H, Guo X, Narisu N, Lan T, Wu Q, Xing Y, et al. Whole-genome sequencing of 175 Mongolians uncovers population-specific genetic architecture and gene flow throughout North and East Asia. Nat Genet. 2018;50:1696–704.
14. Wang H, Yang MA, Wangdue S, Lu H, Chen H, Li L, et al. Human genetic history on the Tibetan plateau in the past 5100 years. Sci Adv. 2023;9:eadd5582.
15. Feng Q, Lu Y, Ni X, Yuan K, Yang Y, Yang X, et al. Genetic history of Xinjiang's Uyghurs suggests bronze age Multiple-Way contacts in Eurasia. Mol Biol Evol. 2017;34:2572–82.
16. Zhang P, Luo H, Li Y, Wang Y, Wang J, Zheng Y, et al. NyuWa genome resource: A deep whole-genome sequencing-based variation profile and reference panel for the Chinese population. Cell Rep. 2021;37:110017.
17. Relling MV, Klein TE, Gammal RS, Whirl-Carrillo M, Hoffman JM, Caudle KE. The clinical pharmacogenetics implementation consortium: 10 years later. Clin Pharmacol Ther. 2020;107:171–5.
18. McQuillan MA, Ranciaro A, Hansen MEB, Fan S, Beggs W, Belay G, et al. Signatures of convergent evolution and natural selection at the alcohol

dehydrogenase gene region are correlated with agriculture in ethnically diverse Africans. Mol Biol Evol. 2022;39:msac183.
19. Sabbagh A, Darlu P, Crouau-Roy B, Poloni ES. Arylamine N-acetyltransferase 2 (NAT2) genetic diversity and traditional subsistence: a worldwide population survey. PLoS ONE. 2011;6:e18507.
20. Hoque KM, Dixon EE, Lewis RM, Allan J, Gamble GD, Phipps-Green AJ, et al. The *ABCG2* Q141K hyperuricemia and gout associated variant illuminates the physiology of human urate excretion. Nat Commun. 2020;11:2767.
21. Caridi G, Lugani F, Angeletti A, Campagnoli M, Galliano M, Minchiotti L. Variations in the human serum albumin gene: molecular and functional aspects. Int J Mol Sci. 2022;23:1159.
22. Khattab A, Haider S, Kumar A, Dhawan S, Alam D, Romero R, et al. Clinical, genetic, and structural basis of congenital adrenal hyperplasia due to 11β-hydroxylase deficiency. Proc Natl Acad Sci U S A. 2017;114:E1933–40.
23. Yan H, Peng B, Liu Y, Xu G, He W, Ren B, et al. Viral entry of hepatitis B and D viruses and bile salts transportation share common molecular determinants on sodium taurocholate cotransporting polypeptide. J Virol. 2014;88:3273–84.
24. Akey JM, Zhang K, Xiong M, Jin L. The effect of single nucleotide polymorphism identification strategies on estimates of linkage disequilibrium. Mol Biol Evol. 2003;20:232–42.
25. Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, et al. An integrated map of structural variation in 2,504 human genomes. Nature. 2015;526:75–81.
26. Chacón-Duque JC, Adhikari K, Fuentes-Guajardo M, Mendoza-Revilla J, Acuña-Alonzo V, Barquera R, et al. Latin Americans show wide-spread converso ancestry and imprint of local native ancestry on physical appearance. Nat Commun. 2018;9:5388.
27. Wahli W, Michalik L. PPARs at the crossroads of lipid signaling and inflammation. Trends Endocrinol Metab. 2012;23:351–63.
28. Fan W, Waizenegger W, Lin CS, Sorrentino V, He MX, Wall CE, et al. PPARδ promotes running endurance by preserving glucose. Cell Metab. 2017;25:1186–93.
29. Vázquez-Carrera M, Wahli W. PPARs as key mediators in the regulation of metabolism and inflammation. Int J Mol Sci. 2022;23:5025.
30. Voight BF, Kudaravalli S, Wen X, Pritchard JK. A map of recent positive selection in the human genome. PLoS Biol. 2006;4:e72.
31. Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, Cotsapas C, et al. Genome-wide detection and characterization of positive selection in human populations. Nature. 2007;449:913–8.
32. CONVERGE consortium. Sparse whole-genome sequencing identifies two loci for major depressive disorder. Nature. 2015;523:588–91.
33. Speidel L, Forest M, Shi S, Myers SR. A method for genome-wide genealogy Estimation for thousands of samples. Nat Genet. 2019;51:1321–29.
34. Kelleher J, Wong Y, Wohns AW, Fadil C, Albers PK, McVean G. Inferring whole-genome histories in large population datasets. Nat Genet. 2019;51:1330–38.
35. Tang L. Genealogy at the genome scale. Nat Methods. 2019;16:1077.
36. Stern AJ, Wilton PR, Nielsen R. An approximate full-likelihood method for inferring selection and allele frequency trajectories from DNA sequence data. PLoS Genet. 2019;15:e1008384.
37. Hejase HA, Mo Z, Campagna L, Siepel A. A Deep-Learning approach for inference of selective sweeps from the ancestral recombination graph. Mol Biol Evol. 2022;39:msab332.
38. Schrider DR, Kern AD. S/HIC: robust identification of soft and hard sweeps using machine learning. PLoS Genet. 2016;12:e1005928.
39. Ben Elsworth M, Lyon T, Alexander Y, Liu P, Matthews J, Hallett, et al. The MRC IEU OpenGWAS data infrastructure. BioRxiv. 2020. https://doi.org/10.1101/2020.08.10.244293. 2020.08.10.244293v1.
40. Ruth KS, Day FR, Tyrrell J, Thompson DJ, Wood AR, Mahajan A, et al. Using human genetics to understand the disease impacts of testosterone in men and women. Nat Med. 2020;26:252–8.
41. Boyle EA, Li YI, Pritchard JK. An expanded view of complex traits: from polygenic to omnigenic. Cell. 2017;169:1177–86.
42. Giambartolomei C, Vukcevic D, Schadt EE, Franke L, Hingorani AD, Wallace C, et al. Bayesian test for colocalization between pairs of genetic association studies using summary statistics. PLoS Genet. 2014;10:e1004383.
43. Kerimov N, Hayhurst JD, Peikova K, Manning JR, Walter P, Kolberg L, et al. A compendium of uniformly processed human gene expression and splicing quantitative trait loci. Nat Genet. 2021;53:1290–99.
44. Voight BF, Adams AM, Frisse LA, Qian Y, Hudson RR, Di Rienzo A. Interrogating multiple aspects of variation in a full resequencing data set to infer human population size changes. Proc Natl Acad Sci U S A. 2005;102:18508–13.
45. Simon A, Coop G. The contribution of gene flow, selection, and genetic drift to five thousand years of human allele frequency change. Proc Natl Acad Sci U S A. 2024;121:e2312377121.
46. Souilmi Y, Tobler R, Johar A, Williams M, Grey ST, Schmidt J, et al. Admixture has obscured signals of historical hard sweeps in humans. Nat Ecol Evol. 2022;6:2003–15.
47. Mathieson I. Human adaptation over the past 40,000 years. Curr Opin Genet Dev. 2020;62:97–104.
48. Hermisson J, Pennings PS. Soft sweeps and beyond: Understanding the patterns and probabilities of selection footprints under rapid adaptation. Methods Ecol Evol. 2017;8:700–16.
49. Messer PW, Petrov DA. Population genomics of rapid adaptation by soft selective sweeps. Trends Ecol Evol. 2013;28:659–69.
50. Schrider DR, Kern AD. Soft sweeps are the dominant mode of adaptation in the human genome. Mol Biol Evol. 2017;34:1863–77.
51. Johri P, Stephan W, Jensen JD. Soft selective sweeps: addressing new definitions, evaluating competing models, and interpreting empirical outliers. PLoS Genet. 2022;18:e1010022.
52. Hernandez RD, Kelley JL, Elyashiv E, Melton SC, Auton A, McVean G, et al. Classic selective sweeps were rare in recent human evolution. Science. 2011;331:920–4.
53. Tobler R, Souilmi Y, Huber CD, Bean N, Turney CSM, Grey ST, et al. The role of genetic selection and Climatic factors in the dispersal of anatomically modern humans out of Africa. Proc Natl Acad Sci U S A. 2023;120:e2213061120.
54. Tan NS, Vázquez-Carrera M, Montagner A, Sng MK, Guillou H, Wahli W. Transcriptional control of physiological and pathological processes by the nuclear receptor PPARβ/δ. Prog Lipid Res. 2016;64:98–122.
55. Gateva V, Sandling JK, Hom G, Taylor KE, Chung SA, Sun X et al. A large-scale replication study identifies *TNIP1*, *PRDM1*, *JAZF1*, *UHRF1BP1* and *IL10* as risk loci for systemic lupus erythematosus. Nat Genet. 2009;41:1228-33.
56. Bentham J, Morris DL, Graham DSC, Pinder CL, Tombleson P, Behrens TW, et al. Genetic association analyses implicate aberrant regulation of innate and adaptive immunity genes in the pathogenesis of systemic lupus erythematosus. Nat Genet. 2015;47:1457–64.
57. Manning AK, Hivert MF, Scott RA, Grimsby JL, Bouatia-Naji N, Chen H, et al. A genome-wide approach accounting for body mass index identifies genetic variants influencing fasting glycemic traits and insulin resistance. Nat Genet. 2012;44:659–69.
58. Tanaka Y, Bi K, Kitamura R, Hong S, Altman Y, Matsumoto A, et al. SWAP-70-like adapter of T cells, an adapter protein that regulates early TCR-initiated signaling in Th2 lineage cells. Immunity. 2003;18:403–14.
59. Bécart S, Charvet C, Canonigo Balancio AJ, De Trez C, Tanaka Y, Duan W, et al. SLAT regulates Th1 and Th2 inflammatory responses by controlling Ca2+/NFAT signaling. J Clin Invest. 2007;117:2164–75.
60. Singleton KL, Gosh M, Dandekar RD, Au-Yeung BB, Ksionda O, Tybulewicz VL, et al. Itk controls the Spatiotemporal organization of T cell activation. Sci Signal. 2011;4:ra66.
61. Yi W, Gupta S, Ricker E, Manni M, Jessberger R, Chinenov Y, et al. The mTORC1-4E-BP-eIF4E axis controls de Novo Bcl6 protein synthesis in T cells and systemic autoimmunity. Nat Commun. 2017;8:254.
62. Serwas NK, Hoeger B, Ardy RC, Stulz SV, Sui Z, Memaran N, et al. Human DEF6 deficiency underlies an immunodeficiency syndrome with systemic autoimmunity and aberrant CTLA-4 homeostasis. Nat Commun. 2019;10:3106.
63. Kuehn HS, Ouyang W, Lo B, Deenick EK, Niemela JE, Avery DT, et al. Immune dysregulation in human subjects with heterozygous germline mutations in CTLA4. Science. 2014;345:1623–27.
64. Lo B, Zhang K, Lu W, Zheng L, Zhang Q, Kanellopoulou C, et al. Patients with LRBA deficiency show CTLA4 loss and immune dysregulation responsive to abatacept therapy. Science. 2015;349:436–40.
65. Sun C, Molineros JE, Looger LL, Zhou XJ, Kim K, Okada Y, et al. High-density genotyping of immune-related loci identifies new SLE risk variants in individuals with Asian ancestry. Nat Genet. 2016;48:323–30.
66. Yuan Z, Zhong Y, Hu H, Zhang W, Wang G. DEF6 has potential to be a biomarker for cancer prognosis: A pan-cancer analysis. Front Oncol. 2023;12:1064376.
67. Siewert-Rocks KM, Kim SS, Yao DW, Shi H, Price AL. Leveraging gene co-regulation to identify gene sets enriched for disease heritability. Am J Hum Genet. 2022;109:393–404.
68. Amariuta T, Siewert-Rocks K, Price AL. Modeling tissue co-regulation estimates tissue-specific contributions to disease. Nat Genet. 2023;55:1503–11.

69. Shang L, Smith JA, Zhou X. Leveraging gene co-expression patterns to infer trait-relevant tissues in genome-wide association studies. PLoS Genet. 2020;16:e1008734.
70. Arvanitis M, Tayeb K, Strober BJ, Battle A. Redefining tissue specificity of genetic regulation of gene expression in the presence of allelic heterogeneity. Am J Hum Genet. 2022;109:223–39.
71. Wallace C. Eliciting priors and relaxing the single causal variant assumption in colocalisation analyses. PLoS Genet. 2020;16(4):e1008720.
72. Broekema RV, Bakker OB, Jonkers IH. A practical view of fine-mapping and gene prioritization in the post-genome-wide association era. Open Biol. 2020;10:190221.
73. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81:559–75.
74. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. Genetics. 2000;155:945–59.
75. Falush D, Stephens M, Pritchard JK. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics. 2003;164:1567–87.
76. Jakobsson M, Rosenberg NA. CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. Bioinformatics. 2007;23:1801–6.
77. Rosenberg NA. Distruct: A program for the graphical display of population structure. Mol Ecol Notes. 2004;4:137–8.
78. Excoffier L, Lischer HE. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and windows. Mol Ecol Resour. 2010;10:564–7.
79. Pickrell JK, Pritchard JK. Inference of population splits and mixtures from genome-wide allele frequency data. PLoS Genet. 2012;8:e1002967.
80. Fitak RR. *OptM*: estimating the optimal number of migration edges on population trees using *Treemix*. Biol Methods Protoc. 2021;6:bpab017.
81. Weir BS, Cockerham CC. Estimating F-Statistics for the analysis of Population-Structure. Evolution. 1984;38:1358–70.
82. Tajima F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics. 1989;123:585–95.
83. Fu YX, Li WH. Statistical tests of neutrality of mutations. Genetics. 1993;133:693–709.
84. Fay JC, Wu CI. Hitchhiking under positive darwinian selection. Genetics. 2000;155:1405–13.
85. Zeng K, Fu YX, Shi S, Wu CI. Statistical tests for detecting positive selection by utilizing high-frequency variants. Genetics. 2006;174:1431–39.
86. Watterson GA. On the number of segregating sites in genetical models without recombination. Theor Popul Biol. 1975;7:256–76.
87. Librado P, Rozas J. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. Bioinformatics. 2009;25:1451–2.
88. Szpiech ZA, Hernandez RD. Selscan: an efficient multithreaded program to perform EHH-based scans for positive selection. Mol Biol Evol. 2014;31(10):2824–7.
89. International HapMap Consortium, Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, et al. A second generation human haplotype map of over 3.1 million SNPs. Nature. 2007;449:851–61.
90. Pickrell JK, Coop G, Novembre J, Kudaravalli S, Li JZ, Absher D, et al. Signals of recent positive selection in a worldwide sample of human populations. Genome Res. 2009;19:826–37.
91. Beleza S, Santos AM, McEvoy B, Alves I, Martinho C, Cameron E, et al. The timing of pigmentation lightening in Europeans. Mol Biol Evol. 2013;30:24–35.
92. Bandelt HJ, Forster P, Röhl A. Median-joining networks for inferring intraspecific phylogenies. Mol Biol Evol. 1999;16:37–48.

Publisher's note